University of Michigan School of Public Health

The University of Michigan Department of Biostatistics Working Paper Series

Year 2004 *Paper* 23

Monotone Constrained Tensor-product B-spline with application to screening studies

Yue Wang* Jeremy Taylor†

*Vaccine Biometric Research, Merk Co. & Inc.

†University of Michigan, jmgt@umich.edu

Monotone Constrained Tensor-product B-spline with application to screening studies

Yue Wang and Jeremy Taylor

Abstract

When different markers are responsive to different aspects of a disease, combination of multiple markers could provide a better screening test for early detection. It is also resonable to assume that the risk of disease changes smoothly as the biomarker values change and the change in risk is monotone with respect to each biomarker. In this paper, we propose a boundary constrained tensor-product B-spline method to estimate the risk of disease by maximizing a penalized likelihood. To choose the optimal amount of smoothing, two scores are proposed which are extensions of the GCV score (O'Sullivan et al. (1986)) and the GACV score (Ziang and Wahba (1996)) to incorporate linear constraints. Simulation studies are carried out to investigate the performance of the proposed estimator and the selection scores. In addidtion, sensitivities and specificities based ona pproximate leave-one-out estimates are proposed to generate more realisitc ROC curves. Data from a pancreatic cancer study is used for illustration.

Monotone Constrained Tensor-product B-spline with application to screening studies

Yue Wang

Vaccine Biometric Research, Merck Co. & Inc.

and

Jeremy M.G. Taylor*

Department of Biostatistics, University of Michigan

SUMMARY. When different markers are responsive to different aspects of a disease, combination of multiple markers could provide a better screening test for early detection. It is also reasonable to assume that the risk of disease changes smoothly as the biomarker values change and the change in risk is monotone with respect to each biomarker. In this paper, we propose a boundary constrained tensor-product B-spline method to estimate the risk of disease by maximizing a penalized likelihood. To choose the optimal amount of smoothing, two scores are proposed which are extensions of the GCV score (O'Sullivan et al. (1986)) and the GACV score (Xiang and Wahba (1996)) to incorporate linear inequality constraints. Simulation studies are carried out to investigate the performance of the proposed estimator and the selection scores. In addition, sensitivities and specificities based on approximate leave-one-out estimates are proposed to generate more realistic ROC curves. Data

* *email:* jmgt@umich.edu

from a pancreatic cancer study is used for illustration.

1. Introduction

With the development of biotechnology, more biomarkers have been identified as associated with different types of cancer. For example, an elevated PSA level is associated with prostate cancer and elevated $CA125$ is associated with ovarian cancer. $CA19$-9 originally found to be expressed in colorectal cancer patients, has also been identified in patients with pancreatic, stomach, and bile duct cancer. These biomarkers are of potential great use for screening diseased subjects. However, each biomarker on its own may not be sensitive or specific enough for a particular type of disease. When different markers measure different biological aspects of a disease, combination of multiple markers is likely to provide a screening test with better performance.

In the medical screening settings, it is also reasonable to assume that the risk of disease is monotone with respect to each biomarker. In other words, subjects with higher (or lower) biomarker values are likely to have higher risk of disease. In this paper, we are interested in combining different markers together for screening while at the same time imposing restrictions that reflect the monotone relationship between the risk of disease and each biomarker.

Let Y denote the group status variable with $Y = 1$ if diseased and $Y = 0$ if non-diseased. $X = (X_1, X_2, ..., X_p)$ denotes the p-by-1 vector for p biomarkers and $\Omega \in \Re^p$ denotes the design space for X. For a screening test, a subspace of Ω will be defined as the "positive region" according to some criteria. A subject is screened "positive" if the values of his markers fall into the positive region and screened "negative" otherwise. Performance of a screening test is usually assessed by its *sensitivity* and *specificity* where *sensitivity* is the probability that a diseased subject is called positive by the test and *specificity* is the probability that a non-diseased subject is called negative

by the test. A test with higher sensitivity and specificity is considered better. By varying the criteria for the positive region, a set of pairs *(specificity, sensitivity)* can be generated and plotted as *sensitivity* versus *(1-specificity)* to give the receiver operating characteristic (ROC) curve. The ROC curve or summary measures derived from the curve are often used to assess the diagnostic ability of the markers.

Su and Liu (1993) provided the solution of the best linear combination of markers, under a multivariate normal assumption. The estimate is best in the sense that the area under the ROC curve is maximized among all linear combinations. In this paper, we are interested in methods to combine different biomarkers without distribution assumptions about the markers or pre-specification of the boundary shape for the positive region. Baker (2000) proposed a class of non-parametric algorithms that combine multiple markers by generalizing the idea of cut-points to positive regions in multiple dimensions. His methods utilize the ordered categorical markers to construct a positive region that optimizes part of the ROC curve. Monotonicity constraints are imposed on the positive region such that higher values of each biomarker imply a greater probability of disease. For biomarkers that are continuous, the algorithms in Baker (2000) requires that the values of each markers to be divided into sub-categories to generate the ordered categorical variables. In this paper, we are interested in methods that generate a positive region using the biomarkers as continuous variables.

Pepe and McIntosh (2002) pointed out that the likelihood ratio function $C(x) = p(x|y = 1)/p(x|y = 0)$ is the optimal score for screening because regions defined by this score maximize the sensitivity at each level of specificity. When the conditional distributions of biomarkers in the two groups are known, we have an explicit expression for the optimal discriminant function determined by $p(x|y = 1)/p(x|y = 0)$. For example, if the biomarkers

follow multivariate normal distribution, the discriminant function is linear in x if the covariance matrices for x in the disease and non-disease groups are proportional to each other and quadratic in x if the covariance matrices are non-proportional. However, if we don't want to make any distribution assumptions about $p(x|y)$, one approach is to derive the discriminant function based on non-parametric estimates of the two conditional densities (Aitchison and Aitken (1976), Hall (1981), Wright and Stander (1997)). Another approach is to use the fact that the likelihood ratio $(p(x|y=1)/p(x|y=0))$ is equivalent to $(Pr(y=1 \mid x)/Pr(y=0 \mid x))$ up to a constant. Thus, to obtain $p(x|y=1)/p(x|y=0)$ is equivalent to estimating $exp(f(x))/(1+exp(f(x)))$ where $f(x) = log(Pr(y=1 \mid x)/Pr(y=0 \mid x))$. Thus to estimate the positive region without distribution assumptions or pre-specification of the boundary shape, instead of estimating the conditional densities separately, we only need to estimate $f(x)$, which can be done through smoothing methods.

Using B-splines is a popular smoothing method in regression (Green and Silverman (1994),de Boor (2001)). A B-spline consists of polynomial pieces joined at certain knots in a way that allows the shape of the spline to be flexible. The function $f(x)$ is a parametric model which lies in the family of smooth functions defined by the span of a set of B-spline basis. This is a very flexible family and thus provides a good approximation to most smooth functions. The degree of smoothness in $f(x)$ is determined by the number and the placement of knots. In this paper, we will use a B-spline based method that also imposes restrictions to ensure a monotone relationship between the risk of disease and each biomarkers. In other words, the probability of disease increases (or decreases) as the value of each biomarker increases, with the other biomarkers set at fixed values. Several different methods have been proposed in the literature to generate a smooth monotone regression func-

tion. One approach is to obtain the estimate through a two stage procedure (Mukerjee (1988), Mammen (1991)). In univariate cases, the estimate f_{SI} is constructed by first smoothing the data using non-parametric regression, followed by isotonisation of this smooth estimate using the pool adjacent violator algorithm. For multivariate x, Mukarjee and Stern (1994) presents a f_{SI} estimator with a simple ad hoc isotonisation procedure for the I step.

Ramsay (1988) proposed to use monotone basis functions to construct a regression spline for a scalar variable. With non-negative coefficients, the regression surface is guaranteed to be monotone. The basis functions they proposed are the integrated B-spline basis. He and Shi (1998) proposed a monotone B-spline smoothing through L_1 optimization in which case monotonicity can be characterized by linear constraints. Kelly and Rice (1990) proposed monotone smoothing based on cubic B-spline for scalar x. By exploiting the property that a B-spline is nondecreasing if the coefficients are nondecreasing, they enforce monotonicity by placing linear inequality constraints on the coefficients. To obtain the estimate, a hybrid approach that includes a roughness penalty, similar to the P-spline approach (Eilers and Marx (1996)), is implemented. Villalobos and Wahba (1987) modified a smoothing spline with the addition of linear inequality constraints to impose monotonicity.

We propose a boundary constrained tensor-product B-spline. It is motivated by generalizing the methods of Kelly and Rice (1990) into high-dimensional scenarios. In addition, through a re-parameterization, we show that a set of simple boundary constraints can ensure monotonicity. It also differs from much of the previous work in that the response variable y is binary rather than continuous.

In section 2, we review B-splines and describe the monotone tensor-product B-spline and the penalized likelihood estimation method. In section

3, generalized cross-validation scores are used to select the smoothing parameter. In section 4, simulation studies are presented. In section 5, we propose a cross-validated version of an ROC curve for a more honest representation of the properties of the estimated discriminant function. Finally, data from a pancreatic cancer study are used for illustration.

2. Monotone Tensor-product B-spline

2.1 *B-spline*

For one dimensional X, to construct a B-spline of degree q, we place $k+1$ interior knots which divide the range, $[x_{min}, x_{max}]$, into k intervals. An additional q knots are placed at each end of the interval. Let $\mathbf{t} = (t_1, \cdots, t_{k+2q+1})$ denote the knot sequence. Then,

$$f(x) = \sum_{j=1}^{k+q} \alpha_j B_{j,q,\mathbf{t}}(x), \text{ for } x \in [x_{min}, x_{max}], \tag{1}$$

where $B_{j,q,\mathbf{t}}(x)$ is the jth B-spline base that can be computed from the recurrent relation (de Boor (2001)):

$$\begin{aligned} B_{j,0,\mathbf{t}}(x) &= \left\{ \begin{array}{ll} 1 & \text{if } t_j \le x < t_{j+1} \\ 0 & \text{otherwise} \end{array} \right\} \\ B_{j,q,\mathbf{t}}(x) &= w_{j,q,\mathbf{t}} B_{j,q-1,\mathbf{t}}(x) + (1 - w_{j+1,q,\mathbf{t}}) B_{j+1,q-1,\mathbf{t}}(x), \quad w_{j,q,\mathbf{t}}(x) = \frac{x - t_j}{t_{j+q-1} - t_j}. \end{aligned}$$

$B_{j,q,\mathbf{t}}(x)$ consists of $q+1$ polynomial pieces, each of degree q. The polynomial pieces join at q inner knots where the derivative up to degree $q-2$ are continuous. $B_{j,q,\mathbf{t}}(x)$ is positive on the domain (t_j, t_{j+q+1}) and is zero everywhere else. At a given value of x, $q+1$ of the basis functions $B_{j,q,\mathbf{t}}(x)$ are nonzero.

2.2 *Boundary constrained B-spline*

The first derivative of $f(x)$ on $[x_{min}, x_{max}]$ is

$$D(\sum_{i=1}^{k+q} \alpha_i B_{i,q,\mathbf{t}})(x) = \sum_{j=2}^{k+q} (q-1) \frac{\alpha_j - \alpha_{j-1}}{t_{j+q-1} - t_j} B_{j,q-1,\mathbf{t}}(x).$$

Since B-spline basis functions are non-negative for all $q \ge 1$, $f'(x)$ is non-negative if the coefficients are nondecreasing. The method Kelly and Rice

(1990) developed to generate a monotone smoothing surface is based on this property. They proposed to fit a regression model such that

$$f(x) = \sum_{i=1}^{k+q} \alpha_i B_{i,q,\mathbf{t}}(x) \text{ with } \alpha_1 \leq \alpha_2 \leq \cdots \alpha_{k+q} \tag{2}$$

Now suppose we use a different parameterization and let $\alpha_i = \sum_{j=1}^{i} \beta_j$, then (2) becomes

$$f(x) = \sum_{j=1}^{k+q} \beta_j (\sum_{i=j}^{k+q} B_{i,q,\mathbf{t}}(x)), \text{ where } \beta_j \geq 0, j = 2, \cdots, k+q. \tag{3}$$

While the linear inequality constrained model (2) and the boundary constrained model (3) are equivalent, computationally, (3) is much easier to handle.

Notice that (3) is close to the formulation suggested by Ramsay (1988), who proposed to use monotone basis functions so that non-negative coefficients will ensure monotonicity. By integrating the B-spline basis, they constructed their monotone basis which have the same form as $(\sum_{i=j}^{k+q} B_{i,q,\mathbf{t}}(x))$ in (3). However, we allow β_1 to be unconstrained so that the regression surface is not forced to be greater or equal to zero.

2.3 *Boundary constrained Tensor-product B-spline*

For a high dimensional situation, we use the tensor-product of one-dimensional B-splines. For example, for $X \in \Re^2$, let $B_{i,q,\mathbf{t}}$ be the ith B-spline of degree q with knot sequence $\mathbf{t} = (t_1, t_2, ..., t_{m1+q+1})$ and $B_{j,p,\mathbf{s}}$ be the jth B-spline of degree p with knot sequence $\mathbf{s} = (s_1, s_2, ..., s_{m2+p+1})$. Then, $f(x_1, x_2)$ is defined to be

$$f(x_1, x_2) = \sum_{i=1}^{m1} \sum_{j=1}^{m2} \alpha_{ij} B_{i,q,\mathbf{t}}(x_1) B_{j,p,\mathbf{s}}(x_2).$$

where the $m1 \cdot m2$ coefficients α_{ij} are parameters to be estimated.

After re-parameterization where $\alpha_{ij} = \sum_{h=1}^{i} \sum_{l=1}^{j} \beta_{hl}$, $f(x_1, x_2)$ becomes

$$f(x_1, x_2) = \sum_{h=1}^{m1} \sum_{l=1}^{m2} \beta_{hl} (\sum_{i=h}^{m1} B_{i,q,\mathbf{t}}(x_1))(\sum_{j=l}^{m2} B_{j,p,\mathbf{s}}(x_2)).$$

The derivatives of f with respect to each marker are

$$\frac{\partial f(x_1,x_2)}{\partial x_1} = \sum_{j=1}^{m2}\sum_{i=2}^{m1}\sum_{l=1}^{j}\frac{\beta_{il}}{(t_{i+q-1}-t_i)}B_{i,q-1,\mathbf{t}}(x_1)B_{j,p,\mathbf{s}}(x_2),$$

and

$$\frac{\partial f(x_1,x_2)}{\partial x_2} = \sum_{i=1}^{m1}\sum_{j=2}^{m2}\sum_{h=1}^{i}\frac{\beta_{hj}}{(s_{j+p-1}-s_j)}B_{i,q,\mathbf{t}}(x_1)B_{j,p-1,\mathbf{s}}(x_2).$$

Hence, $f(x_1,x_2)$ is non-decreasing in both x_1 and x_2 if all of the $\{\beta_{hl}\}_{h=1,l=1}^{m1,m2}$ are non-negative except for β_{11}.

2.4 *Estimation*

The B-spline estimate is obtained by maximizing the likelihood of the data. The smoothness of the estimate is controlled by the number and positioning of the knots. As the number of knots increases, the estimate becomes less smooth. Different methods have been proposed to choose the placement of knots (Friedman and Silverman (1989), Kooperberg and Stone (1992)). Eilers and Marx (1996) proposed a P-spline approach where a relatively large number of knots are placed to ensure flexibility in the shape of the estimates while at the same time preventing over-fitting by the addition of a roughness penalty $J(f)$. In this way the smoothness is controlled by a smoothing parameter instead of by the number of the knots and their position.

For $X \in \Re^2$, the roughness penalty we use is

$$J(f) = \int\int_{[T\times S]}\{(\frac{\partial^2 f}{\partial x_1^2})^2 + 2(\frac{\partial^2 f}{\partial x_1 \partial x_2})^2 + (\frac{\partial^2 f}{\partial x_2^2})^2\}dx_1 dx_2,$$

where $[T\times S]$ represents the region $[t_{q+1}, t_{m_1+1}]\times[s_{k+1}, s_{m_2+1}]$. Then, given the choice of knots and smoothing parameter, for $X \in \Re^2$, we will fit the regression model

$$f(x_1,x_2) = \sum_{h=1}^{m1}\sum_{l=1}^{m2}\beta_{hl}(\sum_{i=h}^{m1}B_{i,q,\mathbf{t}}(x_1))(\sum_{j=l}^{m2}B_{j,p,\mathbf{s}}(x_2))$$

where $\hat{\beta}_{hl}$ are the solution to

$$\begin{aligned} \min_{\beta_{hl}} \quad & -\frac{1}{n}\sum_{i=1}^{n} l_i(y_i, f(x_i)) + \frac{1}{2}\lambda J(f) \\ \text{subject to} \quad & \beta_{hl} \geq 0, \ \ \forall h = 1, \cdots m_1, \ l = 1, \cdots, m_2, \quad \text{except } h = l = 1, \end{aligned}$$

where $l_i(y_i, f(x_i)) = y_i f(x_i) - log(1 + exp(f(x_i)))$.

2.4.1 *Roughness Penalty*

The function $f(x)$ has a closed form expression. Therefore, $J(f)$ can be worked out algebraically. Use subscript j to indicate the pair (j_1, j_2) where $j_1 = 1, \cdots, m_1$, $j_2 = 1, \cdots, m_2$, and $j = 1, \cdots, m_1 m_2$. For the tensor-product B-spline (without re-parameterization)

$$f(x_1, x_2) = \sum_{j_1=1}^{m1}\sum_{j_2=1}^{m2} \alpha_{j_1 j_2} B_{j_1,q,\mathbf{t}}(x_1) B_{j_2,p,\mathbf{s}}(x_2),$$

Green and Silverman (1994) show that $J(f)$ has a quadratic form with respect to the coefficients. Let $B_{j_1,q,\mathbf{t}}(x) = \delta_{j_1}$, $B_{j_2,p,\mathbf{s}}(x) = \epsilon_{j_2}$, and $\alpha = (\alpha_1, \cdots, \alpha_{m_1 m_2})^T$ with $\alpha_j = \alpha_{j_1 j_2}$, $j = 1, \cdots, m_1 \cdot m_2$. Then, $f(x_1, x_2) = \sum \alpha_j \delta_{j_1} \epsilon_{j_2}$. Let

$$\begin{aligned} D^{(s)}_{j_1 k_1} &= \int_T \delta^{(s)}_{j_1}(t)\delta^{(s)}_{k_1}(t)dt, \text{ for } j_1, k_1 = 1, \cdots, m_1, \\ G^{(s)}_{j_2 k_2} &= \int_S \epsilon^{(s)}_{j_2}(u)\epsilon^{(s)}_{k_2}(u)du, \text{ for } j_2, k_2 = 1, \cdots, m_2, \end{aligned}$$

and K be a $m_1 m_2 \times m_1 m_2$ matrix defined by

$$K_{jk} = \sum_{s=0}^{2} \begin{pmatrix} 2 \\ s \end{pmatrix} D^{(s)}_{j_1 k_1} G^{(2-s)}_{j_2 k_2}.$$

Then, the penalty for f over $[T \times S]$ has a quadratic form $J(f) = \alpha^T K \alpha$.

In our approach, $f(x_1, x_2)$ has a different parameterization

$$\begin{aligned} f(x_1, x_2) &= \sum_{j_1=1}^{m1}\sum_{j_2=1}^{m2} \alpha_{j_1 j_2} B_{j_1,q,\mathbf{t}}(x_1) B_{j_2,p,\mathbf{s}}(x_2) \\ &= \sum_{j_1=1}^{m1}\sum_{j_2=1}^{m2} (\sum_{h=1}^{j_1}\sum_{l=1}^{j_2} \beta_{hl}) B_{j_1,q,\mathbf{t}}(x_1) B_{j_2,p,\mathbf{s}}(x_2). \end{aligned}$$

Again using subscript j to indicate the pair (j_1, j_2). Let $\beta = (\beta_1, \cdots, \beta_{m_1 m_2})^T$ with $\beta_j = \beta_{j_1 j_2}$. Then, the penalty on f becomes $J(f) = \beta^T A^T K A \beta$ where A is a $m_1 m_2 \times m_1 m_2$ matrix with 0 or 1 elements such that $A\beta = \alpha$.

2.4.2 *Estimation*

Let Z be a $n \times m_1 m_2$ constant matrix with the ijth element

$$Z_{ij} = (\sum_{l=j_1}^{m1} B_{l,q,\mathbf{t}}(x_{i1}))(\sum_{l=j_2}^{m2} B_{l,p,\mathbf{s}}(x_{i2})), i = 1, \cdots, n, j = 1, \cdots, m_1 m_2.$$

Then $f(x_{i1}, x_{i2}) = \sum_j \beta_j Z_{ij}$ and the log-likelihood for the data with binary outcomes becomes

$$\sum_{i=1}^{n} [y_i f(x_i) - \log(1 + \exp(f(x_i)))] = Y^T Z\beta - C^T I_{n1},$$

where I_{n1} is a $n \times 1$ vector of $1's$. $C = (C_1, \cdots, C_n)^T$ with $C_i = log(1 + exp(f(x_i)))$.

Hence, given a smoothing parameter λ, the monotone tensor-product B-spline estimate $\hat{f}_\lambda = Z\hat{\beta}$ is obtained by

$$\min_\beta \ -\frac{1}{n} Y^T Z\beta + C^T I_{n1} + \frac{1}{2}\lambda \beta^T A^T K A \beta \tag{4}$$

$$\text{subject to} \quad \beta_j \geq 0, \quad j = 2, \cdots, m_1 m_2.$$

An algorithm for boundary constraints based on the trust region method is used to find the solution (Conn et al. (1988)).

3. Generalized Cross-validation scores

The estimate $\hat{f}_\lambda$ is obtained by maximizing the penalized likelihood given a smoothing parameter λ. λ controls the tradeoff between the goodness-of-fit and the smoothness of the estimate. How to choose an appropriate value for λ remains an important issue.

The Cross-validation (CV) method is a common approach which mimics the situation of training and testing samples based on the available data. One will leave out data points (x_i, y_i) one at a time as a testing sample and

obtain the estimate $\hat{f}_\lambda^{(-i)}(x_i)$ based on the remaining $n-1$ points. A good choice of λ is the one such that the prediction based on $\hat{f}_\lambda^{(-i)}(x_i)$ is close to y_i based on goodness-of-fit criteria. To avoid calculating $\hat{f}_\lambda^{(-i)}(x_i)$ by omitting one observation at a time, an approximation for the leave-one-out estimate is usually derived. For a non-Gaussian outcome, O'Sullivan et al. (1986) proposed a Pearson Chi-square based generalized CV score

$$GCV(\lambda) = \frac{1}{n}\frac{\sum_{i=1}^n (y_i - \hat{\mu}_{i,\lambda})^2 / V(\hat{\mu}_i)}{[n - tr(HW)]^2},$$

$\mu_i = E[y_i]$, $H = Z(Z^T W Z + n\lambda\Sigma)^{-1} Z^T$, where Z is a constant matrix with the ith row z_i^T satisfying $f(x_i) = z_i^T \beta$. Σ is the constant matrix such that the roughness penalty term $J(f)$ equals $\beta^T \Sigma \beta$. W is the diagonal matrix with the iith element $\hat{\mu}(x_i)(1 - \hat{\mu}(x_i))$ and $V(\mu_i) = \mu_i(1-\mu_i)$ when the outcome is binary. Xiang and Wahba (1996) proposed a Kullback-Leibler distance based generalized approximate cross-validation score

$$GACV(\lambda) = \frac{1}{n}\sum_{i=1}^n [-y_i \hat{f}_\lambda(x_i) + b(\hat{f}_\lambda(x_i))] + \frac{tr(H)}{n}\frac{\sum_{i=1}^n y_i(y_i - \hat{\mu}_{i,\lambda})}{n - tr(HW)}.$$

When there are monotonicity constraints, existing generalized cross-validation scores may not be appropriate for selecting the optimal λ since they are derived without consideration of all the constraints imposed on the βs. Modifications are needed. Let $C\beta \leq 0$ be the constraints imposed. At the solution, part or all of the constraints will be active constraints, i.e.,have $\tilde{C}\beta = 0$ where $\tilde{C}$ is a subset of the constraints in C. By the Kuhn-Tucker (KT) optimality conditions, the linear inequality constrained optimization is equivalent to a linear equality constrained optimization (Gill et al. (1981), Fletcher (1987)) where the set of equality constraints are the active constraints from the inequality problem. As a result, at the solution, the original linear inequality constrained optimization can be re-written as an unconstrained problem after some transformation to eliminate the equality constraints. However note

that the active constraints are only known at the solution of the problem, but not a priori.

For the monotone tensor-product B-spline methods, assume that there are R constraints active at the solution. Let $\tilde{C}$ be the collection of the R rows from the identity matrix I corresponding to the active constraints. Find the QR decomposition of $\tilde{C}^T$ and let $\tilde{Q}_2$ be the $n \times n-R$ matrix such that

$$\tilde{C}^T = [\tilde{Q}_1 \;\; \tilde{Q}_2] \begin{bmatrix} R_{R\times R} \\ 0_{(n-R)\times R} \end{bmatrix}.$$

Then taking the transformation such that $\beta = \tilde{Q}_2\tilde{\beta}$, the equivalent unconstrained optimization is

$$\min l_\lambda(\tilde{\beta}) \;\; = \;\; -\frac{1}{n}Y^T Z\tilde{Q}_2\tilde{\beta} + \frac{1}{n}B^T I_{n\times 1} + \frac{1}{2}\lambda\tilde{\beta}^T\tilde{Q}_2^T\Sigma\tilde{Q}_2\tilde{\beta}.$$

O'Sullivan et al. (1986) derived the GCV score by a first- order Taylor series expansion of the score function around the limiting penalized likelihood estimate. Xiang and Wahba (1996) derived the GACV score by second-order Taylor series expansion of the likelihood with respect to f. The derivation involves a Σ_f term such that the penalty term $J(f)$ can be written as $f^T\Sigma_f f$ where $f = (f(x_1), \cdots, f(x_n))^T$. We will approximate the leave-one-out estimate $f_\lambda^{(-i)}(x_i)$ by $z_i^T\hat{\beta}^{(-i)}$. $\hat{\beta}^{(-i)}$ is the approximate leave-one-out estimate which is obtained by adapting the approach commonly used to derive the leave-one-out coefficient estimates in generalized linear models (Cook (1982)).

Given a value of λ, we have

$$\hat{f}_\lambda^{(-i)}(x_i) \approx \hat{f}_\lambda(x_i) - \frac{\tilde{h}_{ii}(y_i - \hat{\mu}_i)}{1 - \tilde{h}_{ii}w_{ii}} \tag{5}$$

where $\tilde{h}_{ii}$ is the ith diagonal element of

$$\tilde{H} = Z\tilde{Q}_2(\tilde{Q}_2^T Z^T W Z\tilde{Q}_2 + n\lambda\tilde{Q}_2^T\Sigma\tilde{Q}_2)^{-1}\tilde{Q}_2^T Z^T.$$

Replace $\tilde{h}_{ii}$ with $Tr(\tilde{H})/n$, then the modified Pearson Chi-squared based generalized cross-validation score for our constrained problem is,

$$GCV(\lambda) = \frac{1}{n}\frac{\sum_{i=1}^{n}(y_i - \hat{\mu}_{i,\lambda})^2/\hat{V}_i}{[n - tr(\tilde{H}W)]^2} \tag{6}$$

and the corresponding Kullback-Leibler distance based generalized approximate cross-validation score is

$$GACV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}[-y_i\hat{f}_\lambda(x_i) + b(\hat{f}_\lambda(x_i))] + \frac{tr(\tilde{H})}{n}\frac{\sum_{i=1}^{n} y_i(y_i - \hat{\mu}_{i,\lambda})}{n - tr(\tilde{H}W)}. \tag{7}$$

A second complication of adding monotonicity constraints is that the constraints by themselves provide some degree of smoothing. Hence, the balance between the goodness-of-fit and the smoothness of the estimate is not just controlled by the smoothing parameter λ, but is also restricted by the monotonicity constraints. This means that subject to the monotonicity constraints, when the value of λ is small enough, as it gets smaller, the estimate won't become rougher as in the unconstrained scenarios. Under these circumstance, the approximation for $\hat{f}_\lambda^{(-i)}(x_i)$ by (5) does not work well because even though $\hat{f}_\lambda(x_i)$ doesn't change much, $\frac{\tilde{h}_{ii}(y_i - \hat{\mu}_i)}{1-\tilde{h}_{ii}w_{ii}}$ will become larger as λ gets smaller. As a result, there is an arbitrary over-correction on $\hat{f}_\lambda(x_i)$ to generate the approximate $\hat{f}_\lambda^{(-i)}(x_i)$.

However, notice that the values of $GCV(\lambda)$ and $GACV(\lambda)$ in (6) and (7) increase as λ decreases when $\hat{f}_\lambda$ stays similar. $\hat{\lambda}$ is chosen to be the one that corresponds to the minimum of GCV or GACV scores. Therefore, the modified GCV and GACV scores are expected to perform reasonably well even though the approximation for $\hat{f}_\lambda^{(-i)}(x_i)$ is not valid when λ is small.

4. Simulation Studies

Simulation studies are conducted to investigate the performance of the constrained tensor-product B-spline. We assess the potential gain from the addition of monotonicity constraints and the accuracies of the modified GCV and GACV scores in the selection of λ.

We focus on a two dimensional problem where the true underlying monotone function $f(X)$ is

$$\begin{aligned} f(x_{i1}, x_{i2}) &= 2.5 - 1.2x_{i1}^{-3} - 1.2(1 + exp(-15 + 6x_{i1}))^{-1} \\ &\quad -1.2x_{i2}^{-3} - 1.2(1 + exp(-15 + 6x_{i2}))^{-1} \end{aligned}$$

which has contour plot and surface plot as shown in Fig 1. The binary responses y_is are generated from independent Bernoulli distribution with mean $exp(f(x_i))/(1 + exp(f(x_i)))$ where x_is are sampled from a bivariate normal distribution. Datasets with sample size 50, 100, and 200 are generated. We chose basis functions of degree 3. The number of interior knots considered are 4, 6 and 8 and the knots have equal distance between them. Five hundred simulations are carried out for each scenario.

[Figure 1 about here.]

4.1 *Selection of the smoothing parameter*

To assess the accuracy of the $\hat{\lambda}$ selected by the GCV and $GACV$ scores, we need to define the optimal smoothing parameter. Let f be the true underlying function. Two criteria are chosen based on which an optimal λ will be determined. One is weighted mean square error ($WMSE$)

$$WMSE = \frac{1}{n}\sum_{i=1}^{n} w_i(\hat{f}_\lambda(x_i) - f(x_i))^2,$$

where w_i is the expected Fisher information for f at the design points and equals to $\mu_i(1-\mu_i)$ for binary y. $\mu_i = E[y_i]$. The other is Kullback-Leibler distance (KLD) between $\hat{f}_\lambda$ and the true f

$$\begin{aligned} KL(f, \hat{f}_\lambda) &= \frac{1}{n}\sum_{i=1}^{n} E[log(\frac{p(y_i, f(x_i))}{p(y_i, \hat{f}_\lambda(x_i))})] \\ &= \frac{1}{n}\sum_{i=1}^{n}[\mu_i f(x_i) - b(f(x_i))] - \frac{1}{n}\sum_{i=1}^{n}[\mu_i \hat{f}_\lambda(x_i) - b(\hat{f}_\lambda(x_i))]. \end{aligned}$$

For each simulated dataset, we define the optimal smoothing parameter λ^* as the one that minimizes $WMSE(\lambda)$ or $KLD(\lambda)$ over the range of possible

values for λ. Note that λ^* may differ slightly depending on which criteria is used, $WMSE$ or KLD.

The performance of GCV and GACV are assessed by comparing the $WMSE$ (KLD) associated with the $\hat{\lambda}$ selected by GCV or $GACV$ to the $WMSE$ (KLD) associated with the optimal λ^*. Let $WMSE(\lambda^*)$ ($KLD(\lambda^*)$) be the one corresponding to λ^*, $WMSE(\hat{\lambda}_{GCV})$ ($KLD(\hat{\lambda}_{GCV})$) corresponding to $\hat{\lambda}$ selected using GCV score, and $WMSE(\hat{\lambda}_{GACV})$ ($KLD(\hat{\lambda}_{GACV})$) corresponding to $\hat{\lambda}$ selected using $GACV$ score. Then, relative efficiencies

$$\begin{aligned}
\text{Eff}(\hat{\lambda})^{GCV}_{WMSE} &= WMSE(\lambda^*)/WMSE(\hat{\lambda}_{GCV}) \\
\text{Eff}(\hat{\lambda})^{GCV}_{KL} &= KLD(\lambda^*)/KLD(\hat{\lambda}_{GCV}) \\
\text{Eff}(\hat{\lambda})^{GACV}_{WMSE} &= WMSE(\lambda^*)/WMSE(\hat{\lambda}_{GACV}) \\
\text{Eff}(\hat{\lambda})^{GACV}_{KL} &= KLD(\lambda^*)/KLD(\hat{\lambda}_{GACV})
\end{aligned}$$

are used to evaluate the effectiveness of the two selecting scores. The values of Eff assess how good the CV scores are at giving an estimate of f that is close to the true f, compared to the best you can do. Therefore the values of Eff is expected to be less than 1.0 with numbers near 1.0 indicating successful choice of λ. Figure 2 shows boxplots of the efficiencies for GCV and GACV across 500 simulations in estimating λ for various number of knots. The boxplots suggest that GCV and GACV both perform relatively well in terms of the efficiency with GCV slightly better in the scenarios with larger sample size (n = 200) and GACV slightly better in the scenarios with smaller sample size (n = 50).

[Figure 2 about here.]

4.2 *Efficiency gain with monotonicity constraints*

WMSE and KLD provide general quantitative summaries of how the estimated regression surface differ from the true underlying f and are hence used to assess the efficiency gain obtained by introducing the monotonicity

constraints. In addition, for our purpose, the constrained tensor-product B-spline is used to estimate a monotone smooth discriminant function for screening or diagnostic tests. In these setting, sensitivity and specificity are of particular interest.

The efficiency gain obtained by adding monotonicity constraints is assessed by $\frac{WMSE(\hat{\lambda})_{uncon}-WMSE(\hat{\lambda})_{con}}{WMSE(\hat{\lambda})_{uncon}}$ or $\frac{KLD(\hat{\lambda})_{uncon}-KLD(\hat{\lambda})_{con}}{KLD(\hat{\lambda})_{uncon}}$, where $KLD(\hat{\lambda})_{uncon}$ and $WMSE(\hat{\lambda})_{uncon}$ are the distance measures at the unconstrained solution. In addition, the mean estimated sensitivity given a cutoff point (0.80) for specificity is evaluated for $\hat{f}_\lambda$ with and without constraints, and for the true underlying f in each simulated dataset. Let $D = \{i : y_i = 1, i = 1, \cdots, n\}$ and $\bar{D} = \{i : y_i = 0, i = 1, \cdots, n\}$. Then for each dataset, the sensitivity is calculated by $\sum_{i\in D} I(\hat{f}_\lambda(x_i) > c)/\sum_{i\in D} y_i$ where c is the minimum in the set $\{t : \sum_{i\in \bar{D}} I(\hat{f}_\lambda(x_i) \leq t)/\sum_{i\in \bar{D}}(1-y_i) >= 0.80\}$. The results are shown in table 1.

We were surprised to observe that the efficiency gain with the addition of constraints is significantly larger when GCV is used as the selection score for the smoothing parameter. Further investigations reveal that the GCV score by O'Sullivan et al. (1986) in unconstrained scenarios has lower efficiencies than the modified GCV score (6) in constrained scenarios (results not shown). It has been reported that GCV (in unconstrained scenarios) tends to undersmooth and often has multiple minima (Hastie and Tibshirani (1990)). We also observe (results not shown) that GCV tends to undersmooth in unconstrained scenarios. One possible explanation for why GCV score (6) in the presence of monotonicity constraints behaves much better is that under the constraints, the estimate won't exhibit a too wiggly form even when λ becomes small and hence avoid multiple minima or severe under-smoothing.

In general, table 1 suggests that, the efficiency gain decreases as the sample size becomes larger. The gain in efficiency varies slightly when different

numbers of interior knots are chosen. However, the gain doesn't necessarily increase when a larger number of interior knots are used. Although the true underlying risk of disease is a monotone function of the markers, when the sample size is small, the unconstrained estimate may or may not be monotone. Under these circumstances, adding monotonicity constraints is actually adding additional information to the data. However, when the sample size grows, it is more likely that the unconstrained estimate will be monotone. Hence, it is expected that the efficiency gain will be larger for smaller sample sizes.

Comparison of the average sensitivities obtained from estimates with or without constraints and that obtained from true f in table 1 provides another means of assessing the properties of $\hat{f}_{\hat{\lambda}}$ as a method for detecting disease. The results also suggests that when GACV is used to select λ, adding constraints results in estimates that are closer to the sensitivity for the true f when the sample size gets smaller. The unconstrained estimates are similar as those with constraints when the sample size is large. However, when GCV is used to select λ, the unconstrained estimates don't have good performance even when the sample size is large. With the monotonicity constraints, the estimates are close to the true sensitivities.

4.3 *Comparing to logistic regression*

Logistic regression with main effects only is probably the simplest method to obtain a discriminant function that is smooth and monotone. No monotonicity constraints are needed and the estimate is obtained by maximizing the likelihood. However, f is pre-specified parametrically and the boundary of the positive region is pre-determined to be linear. Table 2 shows some results from comparing the constrained tensor-product spline to logistic regression. The efficiencies gain displayed are the average of $\frac{WMSE(\hat{\lambda})_{logis}-WMSE(\hat{\lambda})_{tpbs}}{WMSE(\hat{\lambda})_{logis}}$ or $\frac{KLD(\hat{\lambda})_{logis}-KLD(\hat{\lambda})_{tpbs}}{KLD(\hat{\lambda})_{logis}}$ across 500 simulations. *tpbs* stands for constrained

tensor-product B-spline. In addition, the mean estimated sensitivity given cutoff point 0.80 for specificity are evaluated for $\hat{f}_{logis}$. The results suggest that under the circumstances that we have studied, the constrained tensor-product B-spline methods perform better than simple logistic regression.

5. Cross-validated ROC curves

For medical practitioners, plotting the ROC curve (based on the estimate $\hat{f}_\lambda$ from observed data) is a popular way to visually assess the performance of a screening or diagnostic test. However, since both the estimation and assessment are based on the same set of data, the plotted ROC curve will be too "rosy". If there is an independent testing sample $(y_j^*, x_j^*), j = 1, \cdots, n_t$, we can estimate the sensitivity, $Pr(\hat{f}_\lambda(x_j^*) > c | y_j^* = 1)$ and specificity, $Pr(\hat{f}_\lambda(x_j^*) \leq c | y_j^* = 0)$. Varying the cutoff value of c, an *honest* ROC curve for the constructed screening test will be obtained. When an independent testing sample is not available, a leave-one-out cross-validation procedure can be used to approximate such an *honest* ROC curve. Let $D = \{i : y_i = 1, i = 1, \cdots, n\}$ and $\bar{D} = \{i : y_i = 0, i = 1, \cdots, n\}$. For the cross-validation ROC curve, we estimate the sensitivity and the specificity by $\sum_{i \in D} I(\hat{f}_\lambda^{(-i)}(x_i) > c) / \sum_{i \in D} y_i$ and $\sum_{i \in \bar{D}} I(\hat{f}_\lambda^{(-i)}(x_i) \leq c) / \sum_{i \in \bar{D}} (1 - y_i)$ where $\hat{f}_\lambda^{(-i)}$ is the estimate with the ith observation omitted. If $\hat{f}_\lambda^{(-i)}, i = 1, \cdots, n$ are all close to $\hat{f}_\lambda$, then the cross-validated ROC curve is a good approximation of the honest ROC curve for $\hat{f}_\lambda$.

In practice, we propose using (5) to approximate the leave-one-out estimates $\hat{f}_\lambda^{(-i)}(x_i), i = 1, \cdots, n$. We also estimated $\hat{f}_\lambda^{(-i)}$ for each i to assess the performance of the approximation by (5). Three simulated datasets with sample size 50,100 and 200 respectively, are randomly selected. The true leave-one-out ROC curve based on $\{\hat{f}_\lambda^{(-i)}\}_{i=1}^n$ that are obtained by omitting one observation at a time are compared to the approximate leave-one-out ROC curve based on (5) along with the plug-in curves which are constructed

based on the original estimate $\hat{f}_\lambda$. Curves corresponding to four different values of the smoothing parameter are shown in figure 3 for sample size 50 and 200. The $\hat{\lambda}$s selected by GACV and GCV scores are 0.01 for all these three simulated dataset.

[Figure 3 about here.]

The plots suggest that, in unconstrained scenarios, the approximation using (5) for $\hat{f}_\lambda^{(-i)}(x_i), i = 1, \cdots, n$ performs well. For small λ, the degree of over-fitting of the plug in estimate $\hat{f}_\lambda$ is substantial. The cross-validated ROC curve provides a more honest presentation. For the constrained scenarios, as discussed in section 3, when λ becomes small enough, the roughness of f will be restricted by the monotonicity constraints imposed and the estimate $\hat{f}_\lambda$ gives a smaller degree of over-fitting. Hence, the plug in ROC curve has only a small degree of optimism associated with it. Under these circumstances, using formula (5) will results in an arbitrary overcorrection and the approximation is not valid. The smaller λ becomes, the larger the magnitude of the overcorrection. Given that $\hat{\lambda} = 0.01$ was selected by GCV or GACV, the approximation of $\hat{f}_\lambda^{(-i)}$ by (5) does give an honest estimate of the ROC curve.

6. Illustration using pancreatic cancer data

We apply the constrained tensor-product B-spline approach to a dataset from a pancreatic cancer study used in Pepe and Thompson (2000). In this study, data from 90 pancreatic cancer patients and 51 control patients with pancreatitis are collected. Two serum markers, CA 125 and CA19-9, were measured on each patient. The constrained tensor-product B-spline approach is used to estimate the regression surface, using 10 interior knots for each of the covariates.

The upper row of figure 4 shows the contour plots of the predicted probability of disease derived from the estimated tensor-product B-spline using

GCV and GACV as the method of choosing λ, respectively. The estimate of λ are 0.005 for GCV and 0.01 for GACV. The estimates of the regression surface using GCV or GACV are very similar. Overall, the data suggests some nonlinear patterns in the risk of disease as a function of the markers, although the deviation from linearity doesn't seem to be dramatic. The ROC curves for each of the two markers separately are shown in the lower left plot in figure 4. These suggest that CA19-9 is better than CA 125 in terms of better prediction for patients disease status. The plug-in ROC curve and the cross-validated ROC curve based on the approximation proposed in section 5 are shown in the lower right plot in figure 4. Compared to the ROC curves corresponding to the single marker, the tests from the combined markers has a slightly higher ROC curve.

[Figure 4 about here.]

7. Discussion

In this paper, we have proposed a tensor-product B-spline based approach to generate a smooth monotone discriminant function for disease screening based on a combination of markers. B-spline is a popular smoothing method in regression. Another popular smoothing approach, which we didn't explore in this paper, is to use a smoothing spline (Villalobos and Wahba (1987)) with monotonicity constraints. The smoothing spline belongs to an infinite space of smooth functions. However, we expect little difference practically between the two approaches since the span of the B-spline basis is in general a rich family that allow good approximations to most smooth functions. In addition, a B-spline is likely to be computationally more efficient.

To incorporate the monotonicity constraints, we propose a re-parameterization such that a set of simple boundary constraints ensure monotonicity. The log-likelihood of the data is maximized with a roughness penalty to obtain the estimate. In practice, it is not necessary to include a penalty term when a

small to moderate number of knots are chosen for the B-spline. If the roughness penalty term is not included, one should be more careful about the positioning of the knots as the smoothness of the estimate is likely to be influenced by their placements. However, we recommend to use the monotone tensor-product B-spline approach with a relatively large number of knots to allow a flexible family of curves and to include a roughness penalty term to reduce the variance.

We have investigated the accuracies of the selection scores and the efficiency gain by adding the monotonicity constraints. The simulation results suggest that the modified GCV and GACV scores perform well and the gain in efficiency by adding monotonicity constraints increases when the sample size becomes smaller.

The proposed method can be extended to scenarios where only part of the regression surface is restricted. The constraints are imposed only on the coefficients for those basis functions that are non-zero within the region where the corresponding regression surface is believed to be monotone.

Although we use cases where X is of two dimension as illustration, the method can be extended to higher dimension. High dimension smoothing is generally a harder problem because the data is likely to be sparse and large sample sizes will be needed to get a reasonable estimate. In these situations, people often resort to more restricted families of smooth functions, such as generalized additive models. For cases where the underlying multivariate regression function is monotone, adding monotonicity constraints provides additional information that may not be obvious from the data. Therefore, the advantage of adding monotonicity constraints could be even greater in high dimensional scenarios. The down side of imposing monotonicity constraints in high dimension smoothing is that computationally it could be slow since the number of constraints needed will be large. Under these circumstances,

the gain in computation efficiency by re-parameterizing the linear inequality constrained problem into a simple boundary constrained B-spline will be greater.

8. Acknowledgment

The authors are grateful to Jeffrey Fessler for stimulating discussions and helpful comments. This work was partially supported by NIH grant CA 86400.

References

Aitchison, J. and Aitken, C. (1976). Multivariate binary discrimination by the kernel method. *Biometrika* **63**, 413–420.

Baker, S. G. (2000). Identifying combinations of cancer markers for further study as triggers of early intervention. *Biometrics* **56**, 1802–1087.

Conn, A., Gould, N. and Toint, P. (1988). Testing a class of methods for solving minimization problems with simple bounds on the variables. *Mathematics of computation* **50**, 399–430.

Cook, R. D. (1982). *Residuals and influence in regression.* Chapman and Hall, New York.

de Boor, C. (2001). *A practical guide to splines.* Springer, New York.

Eilers, P. and Marx, B. (1996). Flexible smoothing with b-splines and penalties. *Statistical Science* **11**, 89–121.

Fletcher, R. (1987). *Practical Methods of Optimization.* John Wiley, Chichester.

Friedman, J. and Silverman, B. (1989). Flexible parsimonious smoothing and additive modeling (with discussion). *Technometrics* **31**, 3–39.

Gill, P. E., Murray, W. and Wright, M. (1981). *Practical Optimization.* Academic Press, London.

Green, P. and Silverman, B. (1994). *Nonparametric Regression and Generalized Linear Models*. Chapman and Hall, New York.

Hall, P. (1981). On nonparametric multivariate binary discrimination. *Biometrika* **68**, 287–294.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*. Chapman and Hall, New York.

He, X. and Shi, P. (1998). Monotone b-spline smoothing. *Journal of the American Statistical Association* **93**, 643–650.

Kelly, C. and Rice, J. (1990). Monotone smoothing with application to dose-response curves and the assessment of synergism. *Biometrics* **46**, 1071–1085.

Kooperberg, C. and Stone, C. (1992). Logspline density estimation for censored data. *J. Comput. Graph. Statist.* **1**, 301–328.

Mammen, E. (1991). Estimating a smooth monotone regression function. *Ann. Statist.* **19**, 724–740.

Mukarjee, H. and Stern, S. (1994). Feasible nonparametric estimation of multiargument monotone functions. *Journal of the American Statistical Association* **89**, 77–80.

Mukerjee, H. (1988). Monotone nonparametric regression. *Ann. Statis.* **16**, 741–750.

O'Sullivan, F., Yandell, B. S. and Raynor, W. (1986). Automatic smoothing of regression functions in generalized linear models. *Journal of the American Statistical Association* **81**, 96–103.

Pepe, M. and McIntosh, M. (2002). Combining several screening tests: optimality of the risk score. *Biometrics* **58**, 657–664.

Pepe, M. and Thompson, M. (2000). Combining diagnostic test results to increase accuracy. *Biostatistics* **1**, 123–140.

Ramsay, J. (1988). Monotone regression splines in action. *Statistical Science*

3, 425–441.

Su, J. Q. and Liu, J. S. (1993). Linear combinations of multiple diagnostic markers. *Journal of the American Statistical Association* **88**, 1350–1355.

Villalobos, M. and Wahba, G. (1987). Inequality-constrained multivariate smoothing splines with application to the estimation of posterior probabilities. *Journal of the American Statistical Association* **82**, 239–248.

Wright, D. and Stander, J. (1997). Nonparametric density estimation and discrimination from images of shapes. *Appl. Statis.* **46**, 365–380.

Xiang, D. and Wahba, G. (1996). A generalized approximate cross validation for smoothing splines with non-gaussian data. *Statistica Sinica* **6**, 675–692.

Table 1

Simulation Results

sample size	grid density	% gain with constraints[a] KLD	% gain with constraints[a] WMSE	empirical sensitivity[a] true p	empirical sensitivity[a] constrained $\hat{p}$	empirical sensitivity[a] unconstrained $\hat{p}$
50	(8,8)	27.33	32.81	0.61	0.62	0.71
	(6,6)	23.48	28.15	0.61	0.62	0.70
	(4,4)	16.81	21.37	0.61	0.61	0.67
100	(8,8)	20.71	24.08	0.59	0.59	0.65
	(6,6)	17.72	20.69	0.59	0.59	0.64
	(4,4)	10.85	12.99	0.59	0.60	0.62
200	(10,10)	11.09	13.56	0.61	0.61	0.65
	(8,8)	10.87	13.11	0.61	0.61	0.65
	(6,6)	8.93	11.02	0.61	0.61	0.64

[a] smoothing parameter selected by GCV

sample size	grid density	% gain with constraints[b] KLD	% gain with constraints[b] WMSE	empirical sensitivity[b] true p	empirical sensitivity[b] constrained $\hat{p}$	empirical sensitivity[b] unconstrained $\hat{p}$
50	(8,8)	6.56	7.44	0.61	0.62	0.63
	(6,6)	8.14	8.64	0.61	0.62	0.63
	(4,4)	8.24	8.98	0.61	0.61	0.63
100	(8,8)	4.44	4.85	0.59	0.59	0.60
	(6,6)	4.15	4.39	0.59	0.59	0.60
	(4,4)	3.98	4.05	0.59	0.59	0.60
200	(10,10)	2.94	3.40	0.61	0.61	0.61
	(8,8)	3.37	3.81	0.61	0.61	0.61
	(6,6)	3.30	3.59	0.61	0.61	0.61

[b] smoothing parameter selected by GACV

Table 2

Comparison of monotone tensor-product B-spline (mtpbs) to simple logistic regression

sample size	number of interior knots	% gain in efficiency[a] KLD	% gain in efficiency[a] WMSE	% gain in efficiency[b] KLD	% gain in efficiency[b] WMSE	empirical sensitivity true p	$\hat{p}_{logistic}$	$\hat{p}^{a}_{mtpbs}$	$\hat{p}^{b}_{mtpbs}$
50	(4,4)	23.92	13.13	26.65	18.66	0.61	0.55	0.61	0.61
100	(8,8)	20.00	15.52	19.34	16.72	0.59	0.56	0.59	0.59
200	(8,8)	20.52	17.69	17.19	15.33	0.61	0.58	0.61	0.61

[a] smoothing parameter selected by GCV
[b] smoothing parameter selected by GACV

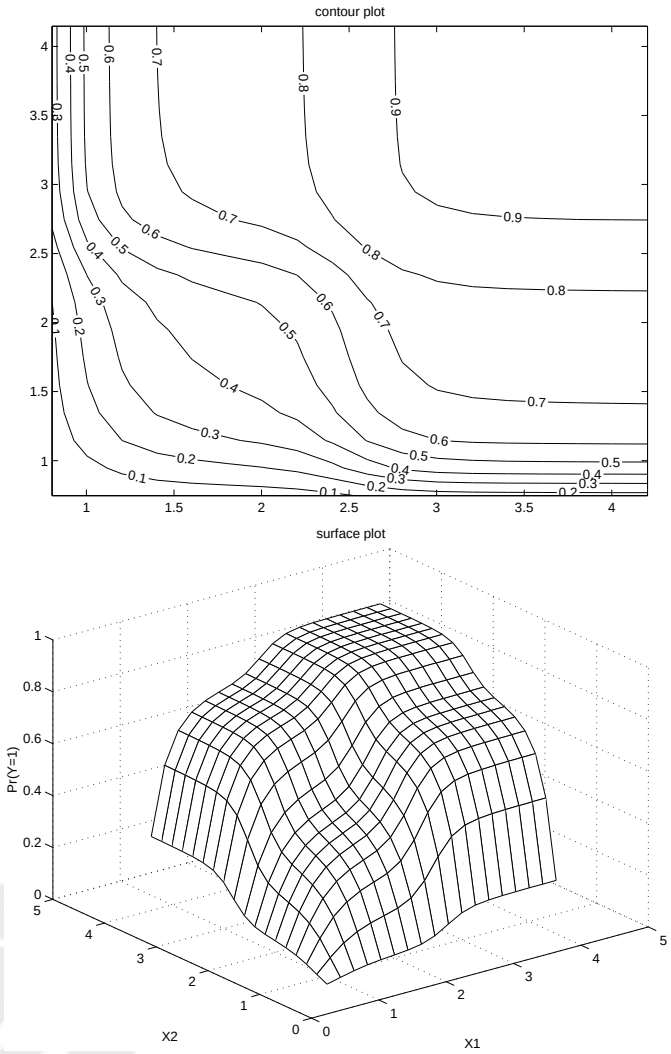

Figure 1. Contour plot and surface plot of the true underlying probability of disease

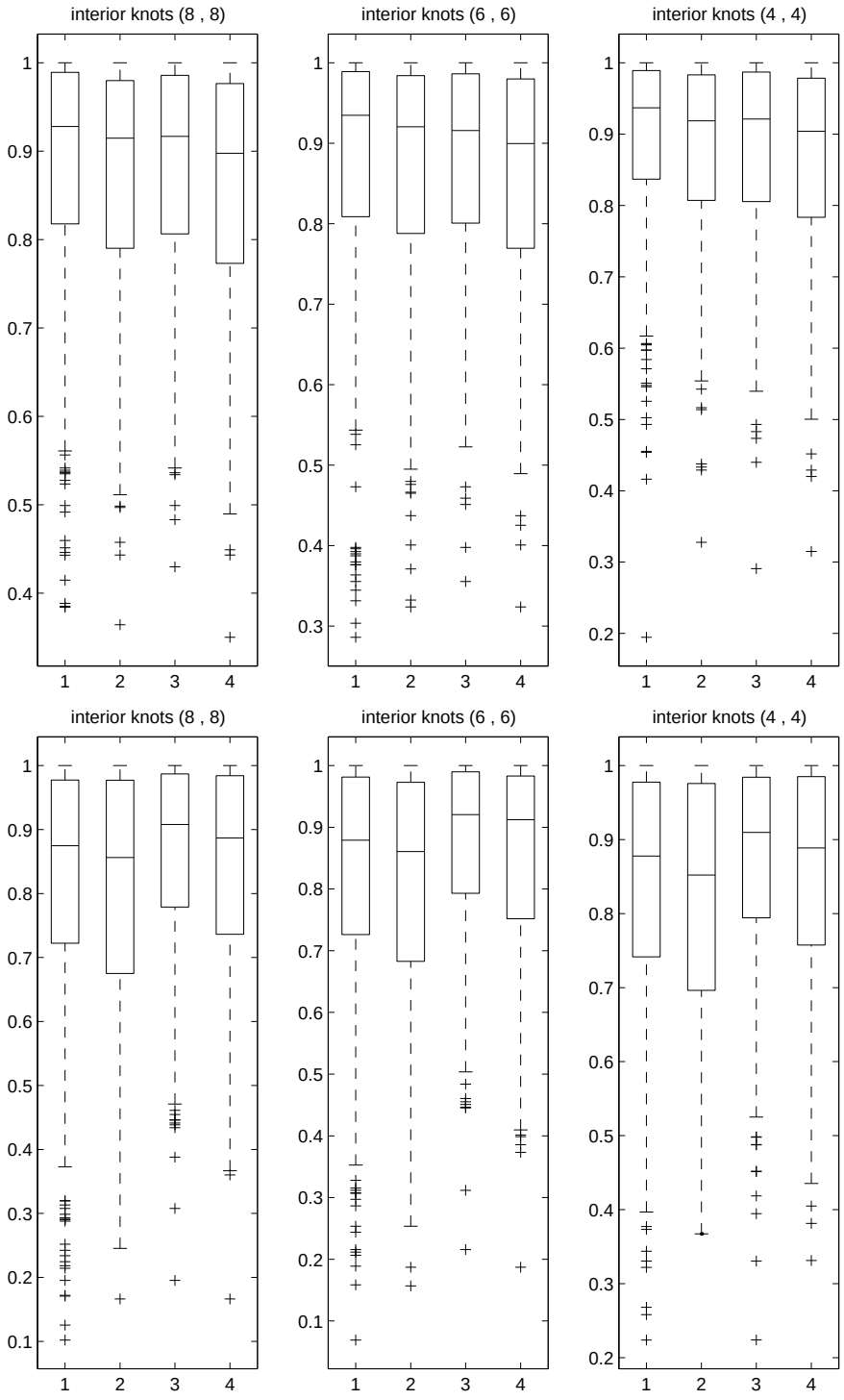

Figure 2. Boxplots of efficiencies of selection scores. Sample size 200(upper row) and 50 (lower row). The vertical scale shows relative efficiency ($WMSE(\lambda^*)/WMSE(\hat{\lambda})$ or $KLD(\lambda^*)/KLD(\hat{\lambda})$. 1, $\hat{\lambda}_{GCV}$ under WMSE criteria; 2, $\hat{\lambda}_{GCV}$ under WMSE criteria; 3, $\hat{\lambda}_{GACV}$ under KLD criteria; 4, $\hat{\lambda}_{GACV}$ under KLD criteria.

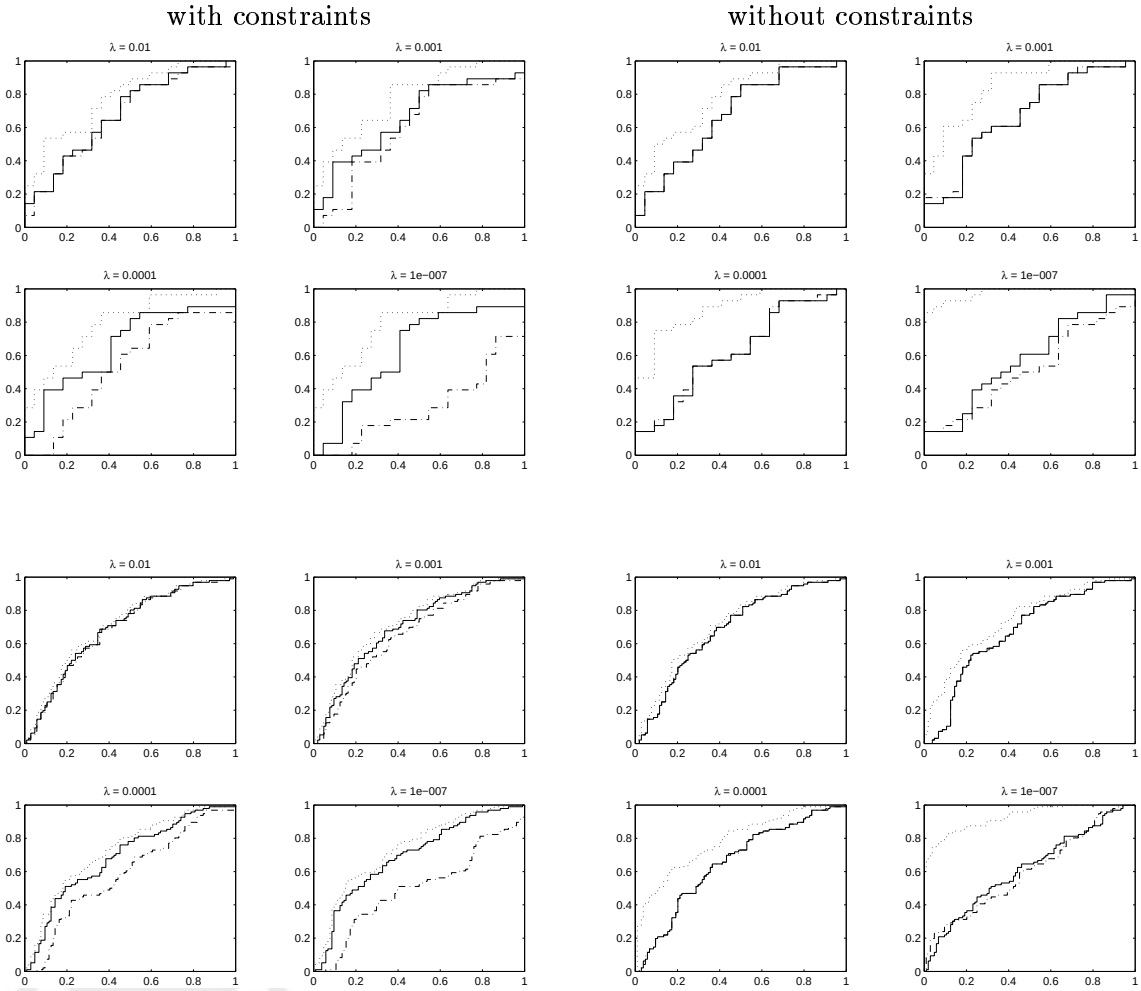

Figure 3. Leave-one-out cross-validated ROC curves (solid line), approximate leave-one-out cross-validated ROC curves (dash-dotted line) and plug-in ROC curves (dotted line). Constrained tensor-product B-splines (left two columns) and unconstrained tensor-product B-splines (right two columns). Sample size 50 (upper two rows) and 200 (lower two rows).

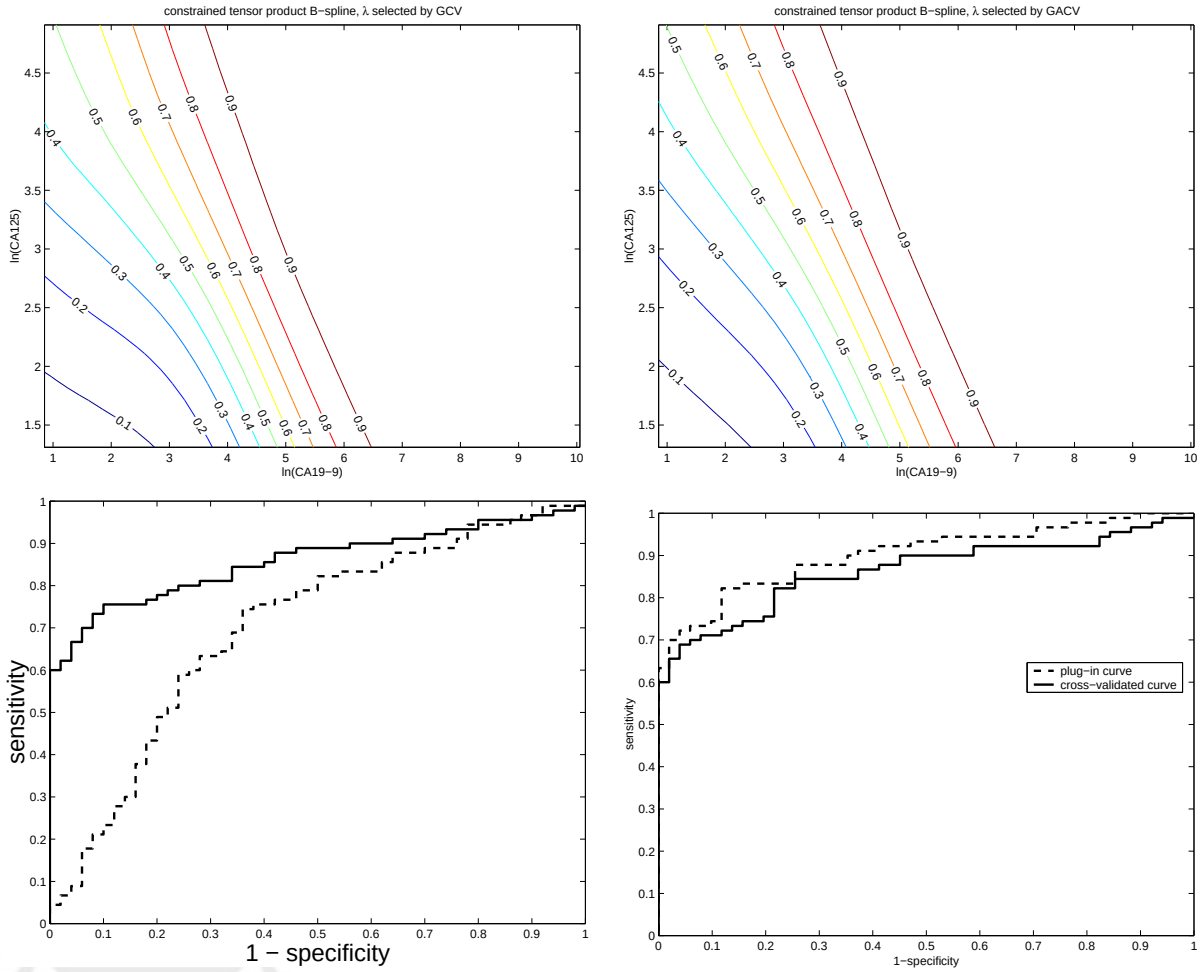

Figure 4. Contour plot of the estimated probability of disease (upper row). Smoothing parameter selected by GCV (upper left, $\hat{\lambda} = 0.005$) and GACV (upper right, $\hat{\lambda} = 0.01$). ROC curves for individual marker (lower left) CA19-9 (solid line) and CA125 (dashed line). ROC curves for combined markers (lower right), Plug-in and cross-validated ROC curve.